Evaluating Self-Report Data Using Psychometric Methods

Ron D. Hays, PhD (hays@rand.org)

February 9, 2005 (3:00-6:00pm)

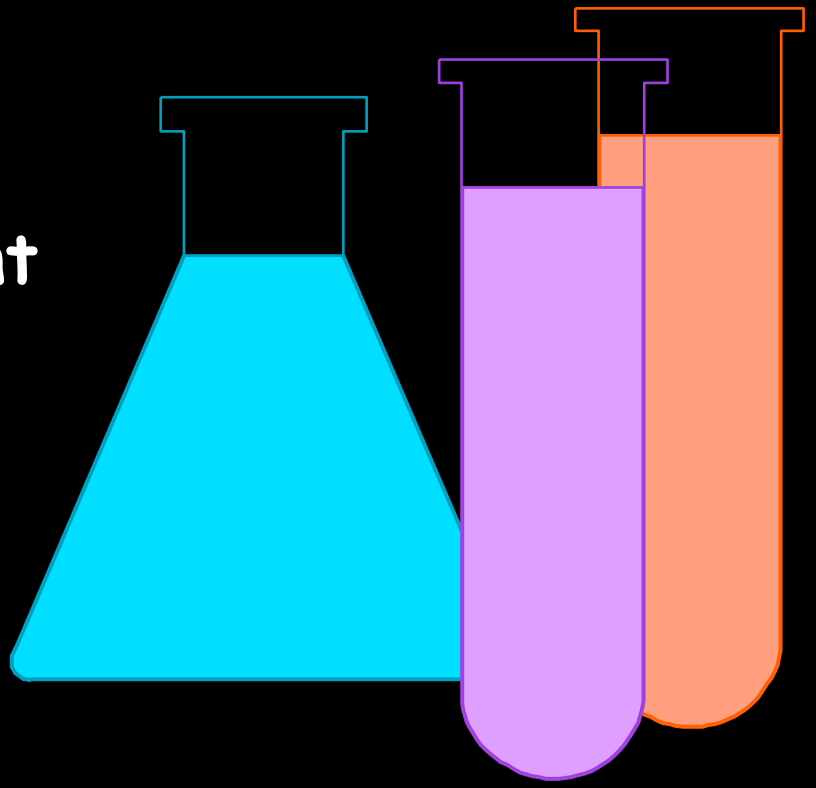
HS 249F

Four Types of Data Collection Errors

- Coverage Error
  Does each person in population have an equal chance of selection?
- Sampling Error
  Are only some members of the population sampled?
- Nonresponse Error
  Do people in the sample who respond differ from those who do not?
- Measurement Error
  Are inaccurate answers given to survey questions?

What's a Good Measure?

- Same person gets same score (reliability)
- Different people get different scores (validity)
- People get scores you expect (validity)
- It is practical to use (feasibility)



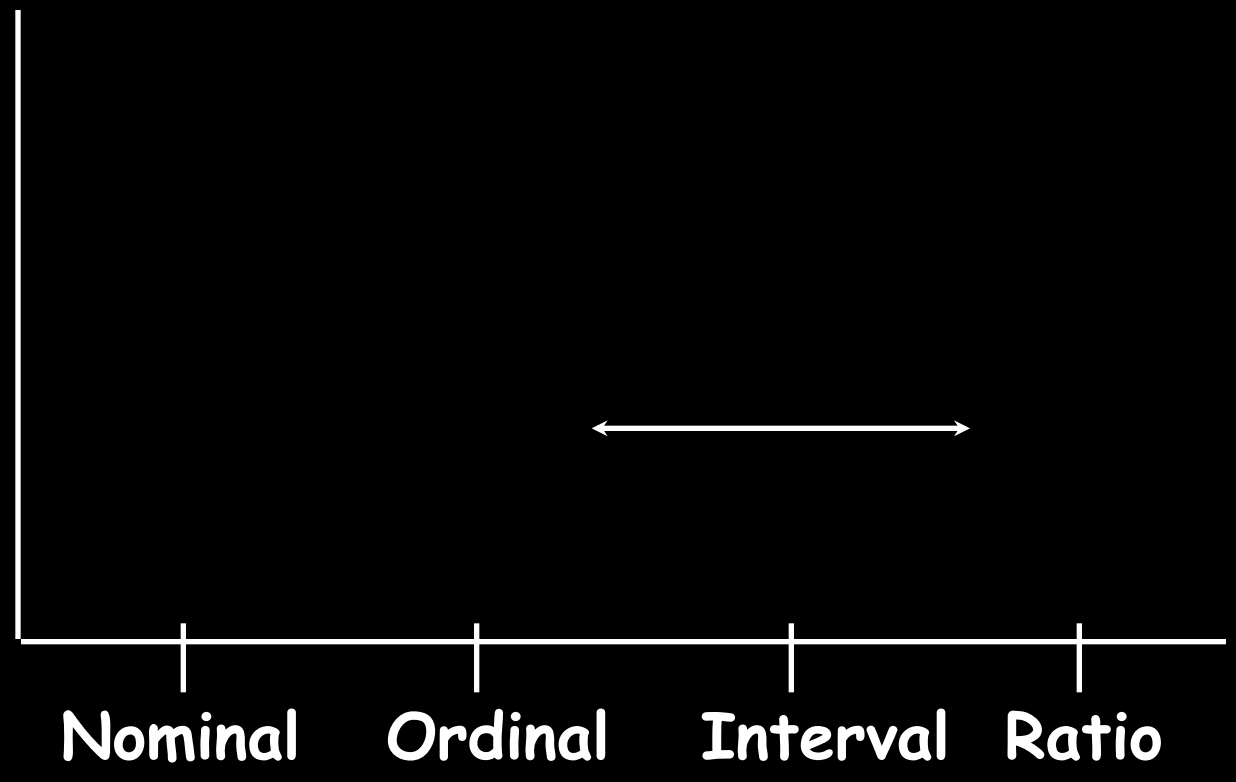
How Are Good Measures Developed?

- Review literature
- Expert input (patients and clinicians)
- Define constructs you are interested in
- Draft items (item generation)
- Pretest
  - Cognitive interviews
  - Field and pilot testing
- Revise and test again
- Translate/harmonize across languages

Scales of Measurement and Their Properties

Type of Scale	Property of Numbers: Rank Order	Equal Interval	Absolute 0
Nominal	No	No	No
Ordinal	Yes	No	No
Interval	Yes	Yes	No
Ratio	Yes	Yes	Yes

Measurement Range for Health Outcome Measures

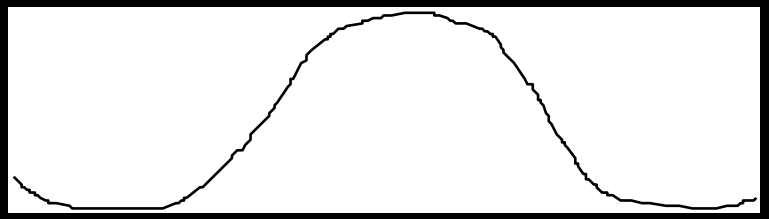


Indicators of Acceptability

- Response rate
- Administration time
- Missing data (item, scale)

Variability

- All scale levels are represented
- Distribution approximates bell-shaped "normal"



Measurement Error

observed = true score + systematic error + random error

(bias)

Flavors of Reliability

- Test-retest (administrations)
- Intra-rater (raters)
- Internal consistency (items)

Test-retest Reliability of MMPI 317-362
r = 0.75

MMPI 362 \ MMPI 317	True	False	
True	169	15	184
False	21	95	116
	190	110	

I am more sensitive than most other people.

Kappa Coefficient of Agreement (Corrects for Chance)

$$\text{kappa} = \frac{(\text{observed} - \text{chance})}{(1 - \text{chance})}$$

Example of Computing KAPPA

		Rater A					
		1	2	3	4	5	Row Sum
Rater B	1	1	1				2
	2		2				2
	3			2			2
	4				2		2
	5					2	2
Column Sum		1	3	2	2	2	10

Example of Computing KAPPA (Continued)

$$P_c = \frac{(1 \times 2) + (3 \times 2) + (2 \times 2) + (2 \times 2) + (2 \times 2)}{(10 \times 10)} = \boxed{0.20}$$

$$P_{obs.} = \frac{9}{10} = \boxed{0.90}$$

$$\text{Kappa} = \frac{0.90 - 0.20}{1 - 0.20} = \boxed{0.87}$$

Guidelines for Interpreting Kappa

Conclusion	Kappa
Poor	< .40
Fair	.40 - .59
Good	.60 - .74
Excellent	> .74

Fleiss (1981)

Conclusion	Kappa
Poor	< 0.0
Slight	.00 - .20
Fair	.21 - .40
Moderate	.41 - .60
Substantial	.61 - .80
Almost perfect	.81 - 1.00

Landis and Koch (1977)

Ratings of Height of Houseplants

Plant		Baseline Height	Follow-up Height	Experimental Condition
A1				
	R1	120	121	1
	R2	118	120	
A2				
	R1	084	085	2
	R2	096	088	
B1				
	R1	107	108	2
	R2	105	104	
B2				
	R1	094	100	1
	R2	097	104	
C1				
	R1	085	088	2
	R2	091	096	

Ratings of Height of Houseplants (Cont.)

Plant		Baseline Height	Follow-up Height	Experimental Condition
C2				
	R1	079	086	1
	R2	078	092	
D1				
	R1	070	076	1
	R2	072	080	
D2				
	R1	054	056	2
	R2	056	060	
E1				
	R1	085	101	1
	R2	097	108	
E2				
	R1	090	084	2
	R2	092	096	

Reliability of Baseline Houseplant Ratings

Ratings of Height of Plants: 10 plants, 2 raters

Baseline Results

Source	DF	SS	MS	F
Plants	9	5658	628.667	35.52
Within	10	177	17.700	
Raters	1	57.8	57.800	
Raters x Plants	9	119.2	13.244	
Total	19	5835		

Sources of Variance in Baseline Houseplant Height

Source	dfs	MS	
Plants (N)	9	628.67	(BMS)
Within	10	17.70	(WMS)
Raters (K)	1	57.80	(JMS)
Raters x Plants	9	13.24	(EMS)
Total	19		

Intraclass Correlation and Reliability

Model	Reliability	Intraclass Correlation
One-Way	$\frac{MS_{BMS} - MS_{WMS}}{MS_{BMS}}$	$\frac{MS_{BMS} - MS_{WMS}}{MS_{BMS} + (K-1)MS_{WMS}}$
Two-Way Fixed	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{EMS} + (K-1)MS_{EMS}}$
Two-Way Random	$\frac{N(MS_{BMS} - MS_{EMS})}{NMS_{BMS} + MS_{JMS} - MS_{EMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (K-1)MS_{EMS} + K(MS_{JMS} - MS_{EMS})/N}$

Summary of Reliability of Plant Ratings

	Baseline		Follow-up	
	R_{TT}	R_{II}	R_{TT}	R_{II}
One-Way Anova	0.97	0.95	0.97	0.94
Two-Way Random Effects	0.97	0.95	0.97	0.94
Two-Way Fixed Effects	0.98	0.96	0.98	0.97

Source	Label	Baseline MS
Plants	BMS	628.667
Within	WMS	17.700
Raters	JMS	57.800
Raters X Plants	EMS	13.244

Cronbach's Alpha

Source	df	SS	MS
Respondents (BMS)	4	11.6	2.9
Items (JMS)	1	0.1	0.1
Resp. x Items (EMS)	4	4.4	1.1
Total	9	16.1	

$$\text{Alpha} = \frac{2.9 - 1.1}{2.9} = \frac{1.8}{2.9} = \boxed{0.62}$$

Alpha by Number of Items and Inter-item Correlations

$$\text{alpha}_{st} = \frac{K\,\bar{r}}{1 + (K - 1)\,\bar{r}}$$

K = number of items in scale

Alpha for Different Numbers of Items and Homogeneity

Number of Items (K)	Average Inter-item Correlation ($\bar{r}$)					
	.0	.2	.4	.6	.8	1.0
2	.000	.333	.572	.750	.889	1.000
4	.000	.500	.727	.857	.941	1.000
6	.000	.600	.800	.900	.960	1.000
8	.000	.666	.842	.924	.970	1.000

Spearman-Brown Prophecy Formula

$$\text{alpha}_y = \left(\frac{N \cdot \text{alpha}_x}{1 + (N - 1) * \text{alpha}_x} \right)$$

N = how much longer scale y is than scale x

Number of Items and Reliability for Three Versions of the Mental Health Inventory (MHI)

Example Spearman-Brown Calculations

MHI-18

$$\frac{18/32\ (0.98)}{(1+(18/32\ -1)*0.98}$$

= 0.55125/0.57125 = 0.96

Reliability Minimum Standards

- 0.70 or above (for group comparisons)
- 0.90 or higher (for individual assessment)
  - SEM = SD (1- reliability)$^{1/2}$

Reliability of a Composite Score

$$\text{Mosier} = 1 - \frac{\Sigma(w_j^2)(S_j^2) - \Sigma(w_j^2)(S_j^2)(\alpha_j)}{\Sigma(w_j^2)(S_j^2) + 2\Sigma(w_j)(w_K)(S_j)(S_K)(r_{jK})}$$

w_j = weight given to component J

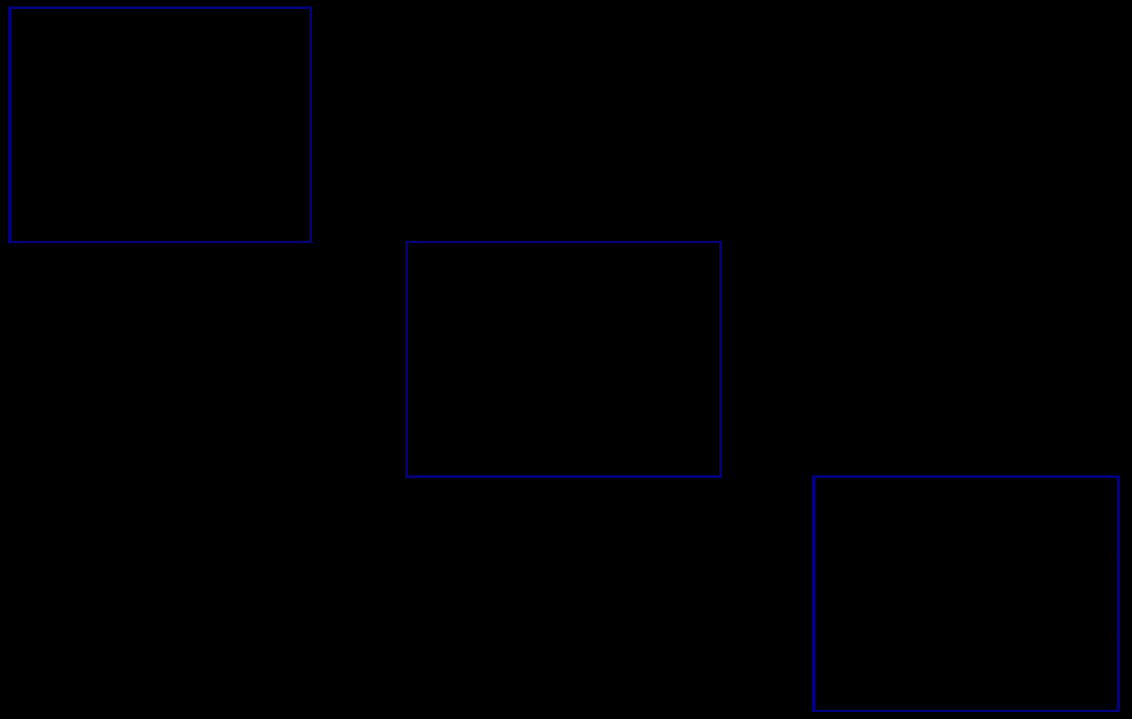
w_K = weight given to component K

S_j = standard deviation of J

α_j = reliability of J

r_{jK} = correlation between J and K

Hypothetical Multitrait/Multi-Item Correlation Matrix



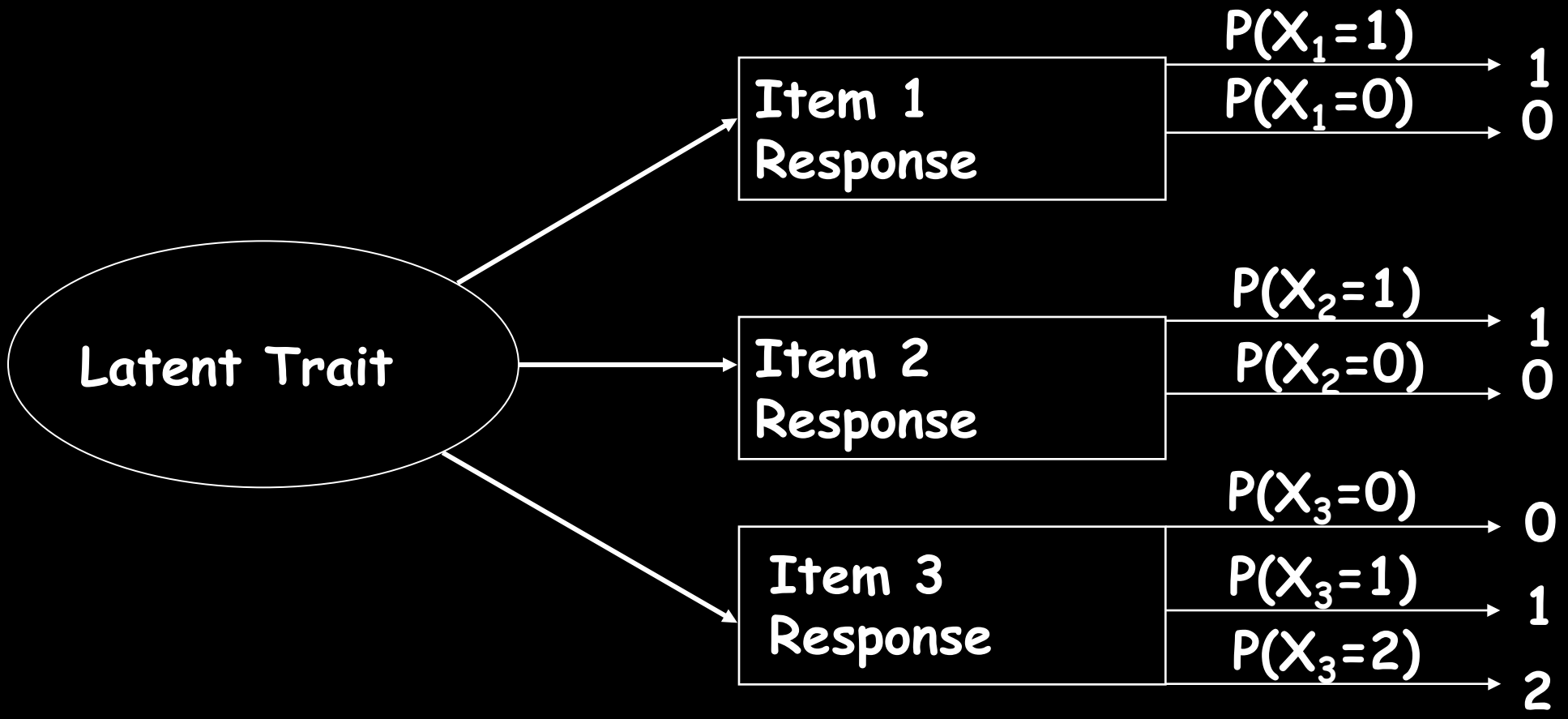
Multitrait/Multi-Item Correlation Matrix for Patient Satisfaction Ratings

	Technical	Interpersonal	Communication	Financial
Technical				
1	0.66*	0.63†	0.67†	0.28
2	0.55*	0.54†	0.50†	0.25
3	0.48*	0.41	0.44†	0.26
4	0.59*	0.53	0.56†	0.26
5	0.55*	0.60†	0.56†	0.16
6	0.59*	0.58†	0.57†	0.23
Interpersonal				
1	0.58	0.68*	0.63†	0.24
2	0.59†	0.58*	0.61†	0.18
3	0.62†	0.65*	0.67†	0.19
4	0.53†	0.57*	0.60†	0.32
5	0.54	0.62*	0.58†	0.18
6	0.48†	0.48*	0.46†	0.24

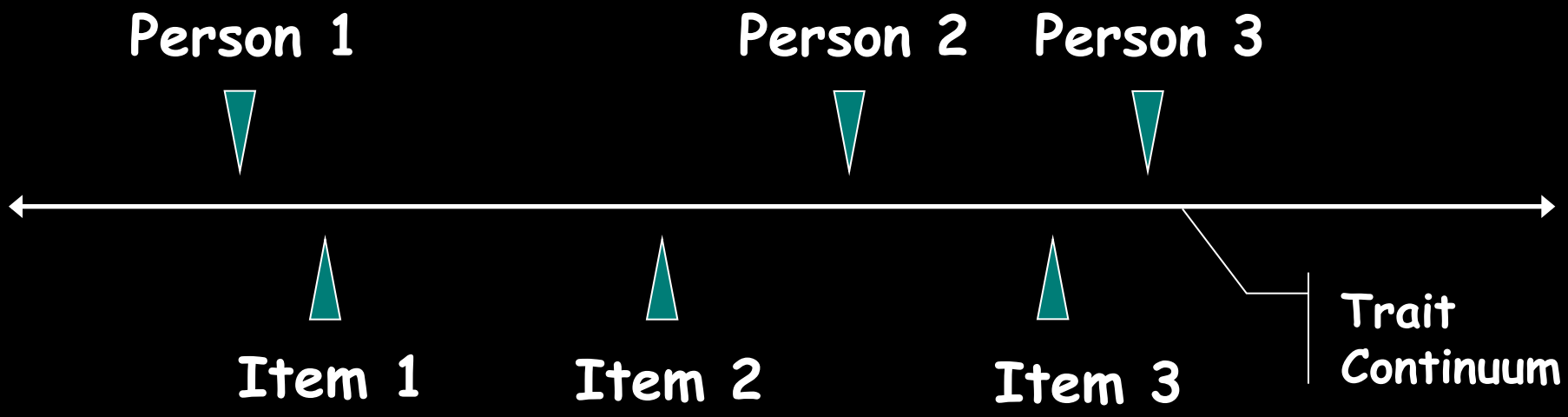
Note – Standard error of correlation is 0.03. Technical = satisfaction with technical quality. Interpersonal = satisfaction with the interpersonal aspects. Communication = satisfaction with communication. Financial = satisfaction with financial arrangements. *Item-scale correlations for hypothesized scales (corrected for item overlap). †Correlation within two standard errors of the correlation of the item with its hypothesized scale.

IRT

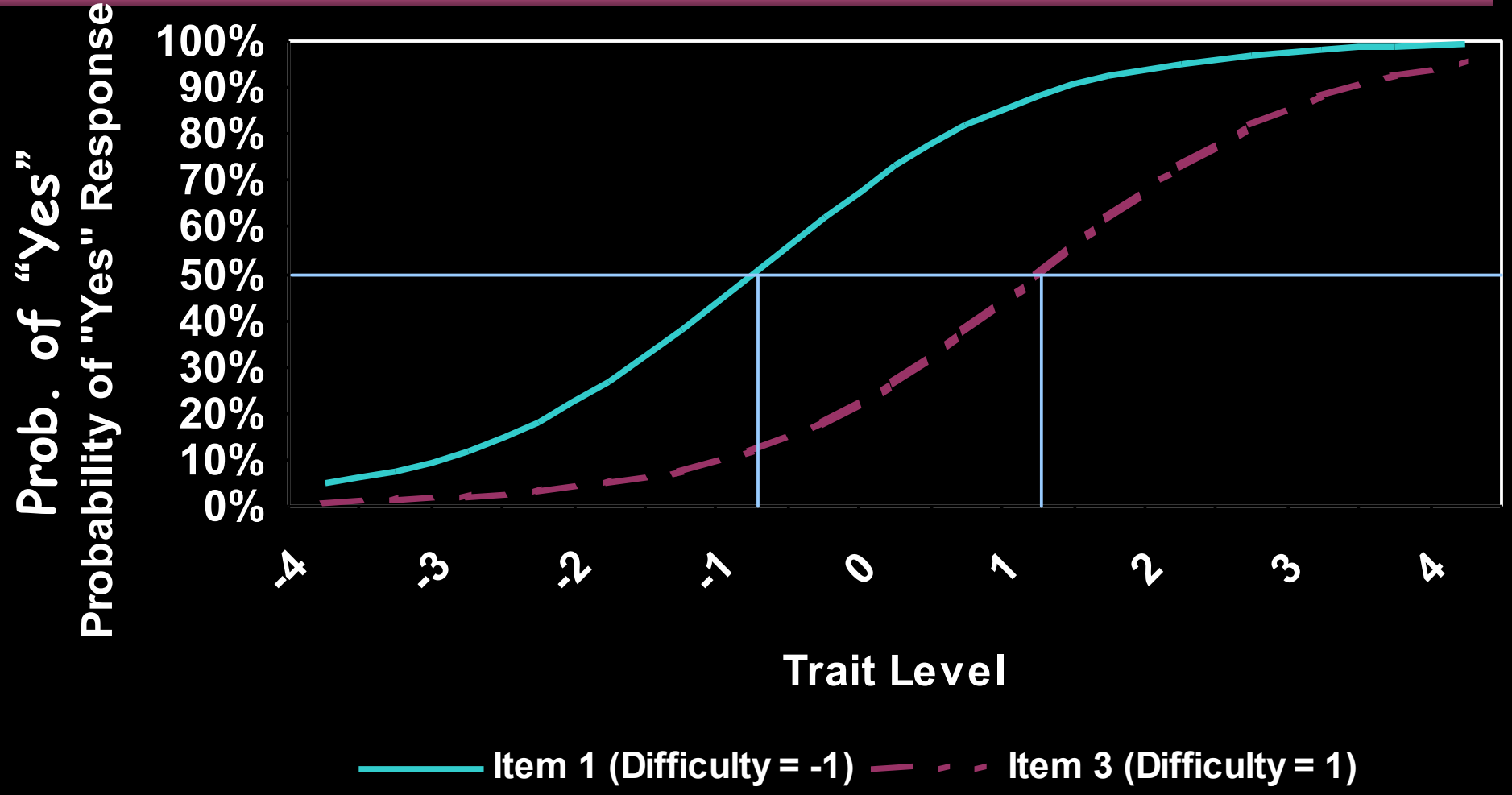
Latent Trait and Item Responses



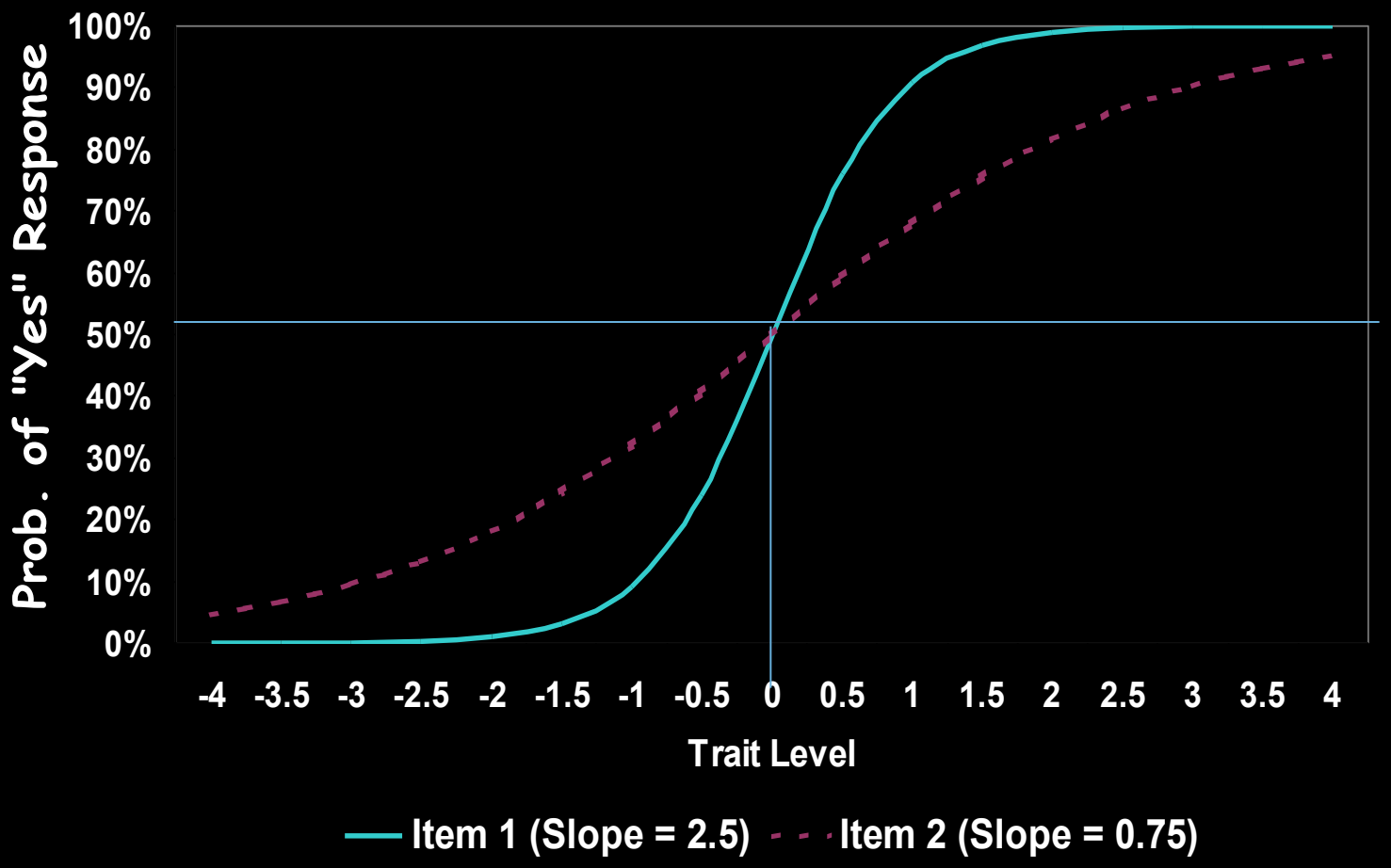
Item Responses and Trait Levels



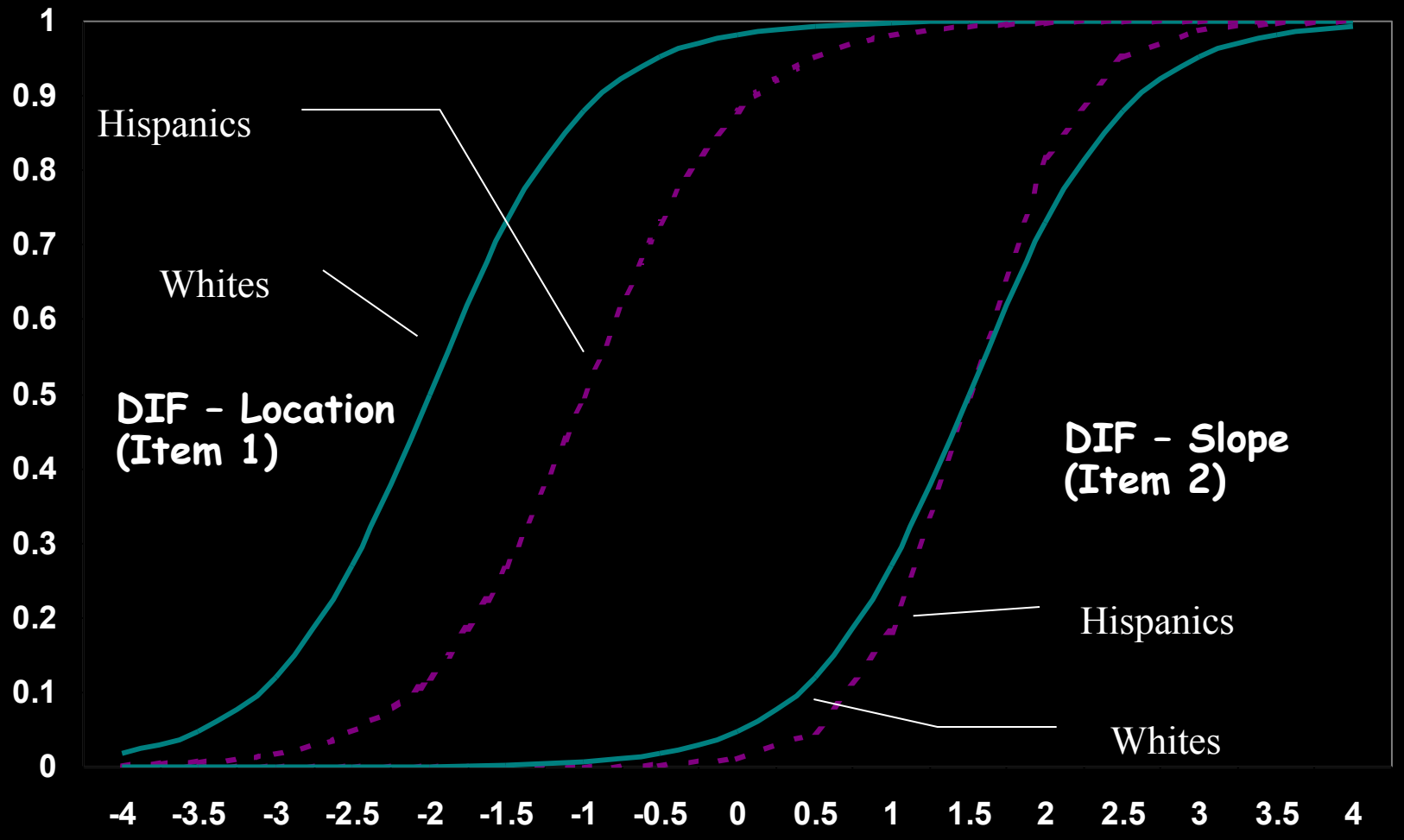
Item Characteristic Curves (1-Parameter Model)



Item Characteristic Curves (2-Parameter Model)



Dichotomous Items Showing DIF
(2-Parameter Model)



Forms of Validity

- Content, Criterion
- Construct Validity

Construct Validity

- Does measure relate to other measures in ways consistent with hypotheses?
- Responsiveness to change

Relative Validity Analyses

- Form of "known groups" validity
- Relative sensitivity of measure to important clinical differences

Relative Validity Example

	Severity of Heart Disease				
	None	Mild	Severe	F-ratio	Relative Validity
Scale #1	91	90	87	2	---
Scale #2	88	78	74	10	5
Scale #3	95	87	77	20	10

Responsiveness to Change and Minimally Important Difference

- HRQOL measures should be responsive to interventions that changes HRQOL

- Evaluating responsiveness requires assessment of HRQOL
  - pre-post intervention of known efficacy
  - at two times in tandem with gold standard

Two Essential Elements

- External indicator of change (Anchors)
  - mean change in HRQOL scores among people who have a "minimal" change in HRQOL.
- Amount of HRQOL change

External Indicator of Change (A)

- **Overall has there been any change in your asthma since the beginning of the study?**

  Much improved; Moderately improved; Minimally improved

  No change

  Much worse; Moderately worse; Minimally worse

External Indicator of Change (B)

Rate your overall condition. This rating should encompass factors such as social activities, performance at work or school, seizures, alertness, and functional capacity; that is, your overall quality of life.

7 response categories; ranging from no impairment to extremely severe impairment

External Indicator of Change (C)

- "changed" group = seizure free (100% reduction in seizure frequency)

- "unchanged" group = <50% change in seizure frequency

Responsiveness Indices

(1) Effect size (ES) = D/SD

(2) Standardized Response Mean (SRM) = D/SD†

(3) Guyatt responsiveness statistic (RS) = D/SD‡

D = raw score change in "changed" group;

SD = baseline SD;

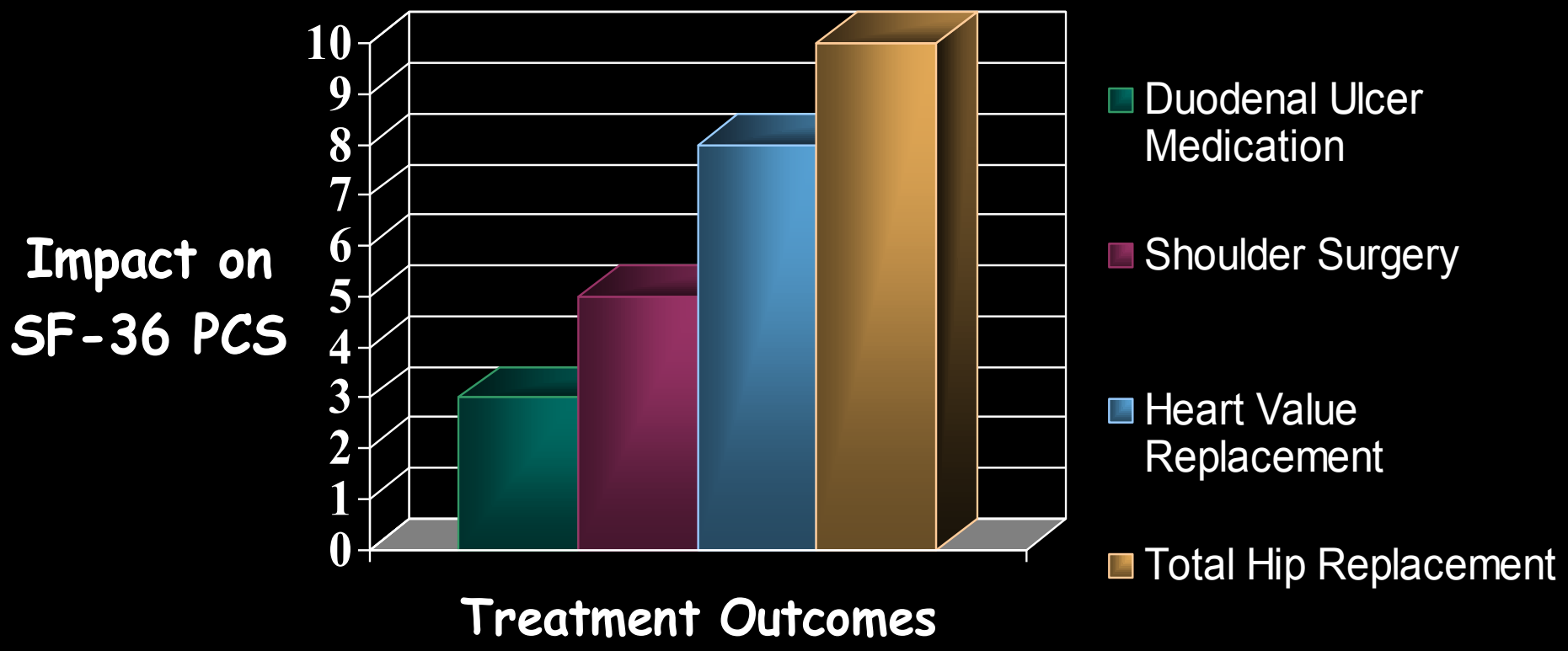
SD† = SD of D;

SD‡ = SD of D among "unchanged"

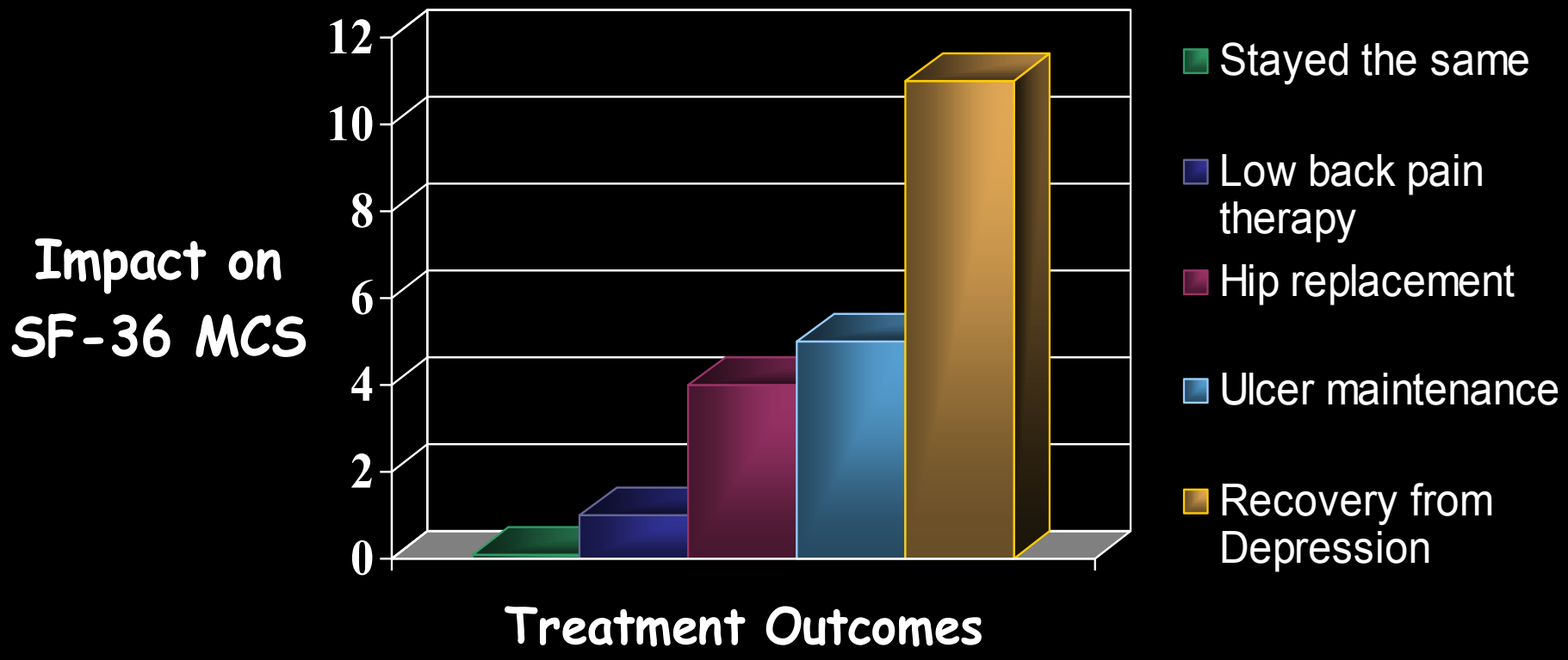
Effect Size Benchmarks

- Small: 0.20->0.49
- Moderate: 0.50->0.79
- Large: 0.80 or above

Treatment Impact on PCS



Treatment Impact on MCS



Individual Change

- Interest in knowing how many patients benefit from group intervention
- Tracking progress on individual patients

Methods

- 54 patients
- Average age = 56; 84% white; 58% female
- Self-administered SF-36 version 2 at baseline and about at end of therapy (about 6 weeks later).

SF-36 Version 2

- Physical functioning (10 items)
- Role limitations/physical (4 items)
- Pain (2 items)
- General health perceptions (5 items)
- Social functioning (2 items)
- Energy/fatigue (4 items)
- Role limitations/emotional (3 items)
- Emotional well-being (5 items)

Scoring the SF-36

- Average or sum all items in the same scale.

- Transform average or sum to
  - 0 (worse) to 100 (best) possible range
  - z-score (mean = 0, SD = 1)
  - T-score (mean = 50, SD = 10)
    - ✓ T-score = 50 + (z-score * 10)

t-test for within group change

- $X_D/(SD_d/n^{1/2})$

X_D = is mean difference, SD_d = standard deviation of difference

Formulas for Significance of Individual Change

Standard error of measurement (SEM)	$SD_b * (1- \text{reliability})^{1/2}$
Standard error of prediction (SEp)	$SD_b * (1- \text{reliability}^2)^{1/2}$
SEM CI around Time 1 score	Time 1 +- 2 SEM
SEp CI around Time 1 score	Time 1 +- 2 SEp

SD_b = standard deviation at baseline

Estimated True Score for Score of 60

- Mean + reliability (score – mean)
- 50 + 0.90 (60 – 50) = 59

Reliable Change Index

- $(X_2 - X_1)$/ (SEM * SQRT [2])

Formulas for Significance of Individual Change

SEM 95% CI	$1.96 * SD_b * (1- \text{reliability})^{1/2}$
SEp 90% CI	$1.64* SD_b * (1- \text{reliability}^2)^{1/2}$
SEp 95% CI	$1.96* SD_b * (1- \text{reliability}^2)^{1/2}$
Estimated true score	Mean + reliability (score – mean)
Reliable change index	$X_2-X_1/$

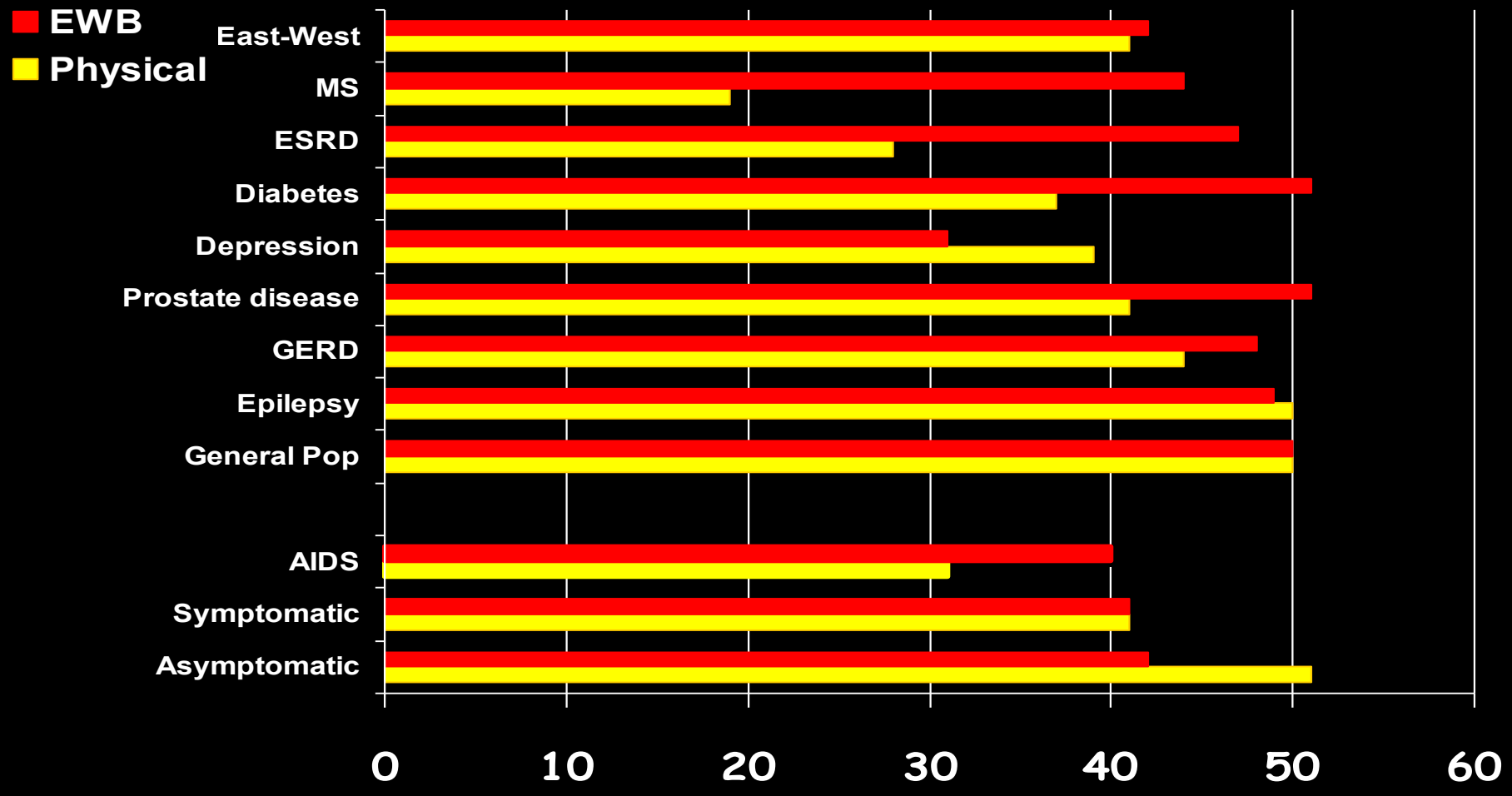
SD_b = standard deviation at baseline

Minimum Delta for Individual Significance

- SEM: > 1.96 SEM
- RCI: > 1.96 * SQRT (2) * SEM

- SEp-90: > 1.64 SEp; SEp-95: > 1.96 SEp

Physical Functioning and Emotional Well-Being at Baseline for 54 Patients at UCLA-Center for East West Medicine



RAND HEALTH

Hays et al. (2000), American Journal of Medicine

Change in SF-36 Scores Over Time

Significance of Group Change

	Delta	t-test	prob.
PF-10	1.7	2.38	.0208
RP-4	4.1	3.81	.0004
BP-2	3.6	2.59	.0125
GH-5	2.4	2.86	.0061
EN-4	5.1	4.33	.0001
SF-2	4.7	3.51	.0009
RE-3	1.5	0.96	.3400 <-
EWB-5	4.3	3.20	.0023
PCS	2.8	3.23	.0021
MCS	3.9	2.82	.0067

Amount of Change in Observed Score Needed for Significant Change

	SEM	90% SEp	95% SEp	RCI	Effect size
PF-10	5.9	6.9	8.2	8.4	0.47-0.67
RP-4	6.0	6.9	8.3	8.4	0.52-0.72
BP-2	7.4	8.4	10.1	10.4	0.72-1.01
GH-5	9.2	10.4	12.5	13.0	0.80-1.13
EN-4	9.0	10.1	12.0	12.8	0.94-1.33
SF-2	9.8	11.1	13.3	13.8	0.76-1.07
RE-3	6.8	8.0	9.5	9.7	0.50-0.71
EWB-5	9.5	10.6	12.7	13.4	0.90-1.26
PCS	5.0	5.9	7.0	7.1	0.43-0.62
MCS	6.9	8.0	9.5	9.7	0.52-0.73

Proportion of 54 Cases Declining Significantly

	SEM	SEp 90	SEp 95	RCI
PF-10	9%	7%	2%	2%
RP-4	7%	6%	2%	2%
BP-2	17%	11%	9%	7%
GH-5	4%	0%	0%	0%
EN-4	4%	4%	2%	2%
SF-2	13%	11%	6%	4%
RE-3	19%	19%	15%	15%
EWB-5	9%	6%	6%	4%
PCS	7%	7%	7%	7%
MCS	13%	11%	11%	11%

Proportion of 54 Cases Improving Significantly

	SEM	SEp 90	SEp 95	RCI
PF-10	19%	15%	13%	13%
RP-4	35%	31%	30%	31%
BP-2	31%	28%	24%	22%
GH-5	9%	7%	7%	7%
EN-4	24%	17%	11%	9%
SF-2	30%	20%	17%	17%
RE-3	24%	19%	15%	15%
EWB-5	26%	20%	19%	19%
PCS	33%	30%	24%	24%
MCS	37%	30%	22%	22%

% Improved - % Declined

	SEM	SEp 90	SEp 95	RCI
PF-10	10%	8%	11%	11%
RP-4	28%	25%	28%	29%
BP-2	14%	17%	15%	15%
GH-5	5%	7%	7%	7%
EN-4	11%	13%	9%	7%
SF-2	11%	9%	11%	13%
RE-3	5%	0%	0%	0%
EWB-5	19%	14%	13%	15%
PCS	26%	23%	17%	17%
MCS	24%	19%	11%	11%